確率・統計基礎: KL ダイバージェンスの性質

森 立平

Kullback–Leibler divergence

Definition (Kullback–Leibler divergence)

$D \colon \mathcal{P} \times \mathcal{P} \to \mathbb{R}$ を以下で定義する

$$D(q \parallel p) \coloneqq \sum_{k=1}^{m} q_k \log \frac{q_k}{p_k}.$$

ただし、$0 \log 0 = 0 \log \frac{0}{0} = 0$ とし、$q \log \frac{q}{0}$ は $q > 0$ について $+\infty$ とする。

サノフの定理

Theorem (サノフの定理)

任意の $\Gamma \subseteq \mathcal{P}$ について

$$\frac{1}{N}\log \Pr\left(\widehat{P}_{\mathbf{x}} \in \Gamma\right) \leq -\inf_{q\in\Gamma} D(q \,\|\, p) + \frac{m\log(N+1)}{N} \quad \textit{for any } N \in \mathbb{Z}_{>0}$$

$$\liminf_{N\to\infty} \frac{1}{N}\log \Pr\left(\widehat{P}_{\mathbf{x}} \in \Gamma\right) \geq -\inf_{q\in\Gamma^\circ} D(q \,\|\, p).$$

ただし、Γ° は Γ の内点の集合とする。

Corollary

$\Gamma \subseteq \mathcal{P}$ について $\Gamma \subseteq \overline{\Gamma^\circ}$ のとき、

$$\lim_{N\to\infty} \frac{1}{N}\log \Pr\left(\widehat{P}_{\mathbf{x}} \in \Gamma\right) = -\min_{q\in\overline{\Gamma}} D(q \,\|\, p).$$

ただし、$\overline{\Gamma}$ は Γ の閉包とする。

KL divergence の性質

1. $D(q \parallel p) \geq 0$ で等号は $q = p$ のときのみ成り立つ。

2. 凸関数である。$D(\lambda q + (1-\lambda)q' \parallel \lambda p + (1-\lambda)p') \leq \lambda D(q \parallel p) + (1-\lambda)D(q' \parallel p')$.

イェンセンの不等式

Lemma (イェンセンの不等式)

凸関数 $f : \mathbb{R} \to \mathbb{R}$ と確率変数 $X : \Omega \to \mathbb{R}$ について、

$$\mathbb{E}[f(X)] \geq f(\mathbb{E}[X]).$$

また、f が狭義凸のとき等号は X が決定的 $(\Pr(X = \mathbb{E}[X]) = 1)$ な場合にのみ成り立つ。

Proof.

$f(x)$ の $x = \mathbb{E}[X]$ における "接線" を

$$y = a(x - \mathbb{E}[X]) + f(\mathbb{E}[X])$$

とおくと、f の凸性より接線は f より下側にある。つまり

$$f(x) \geq a(x - \mathbb{E}[X]) + f(\mathbb{E}[X]) \qquad \forall x \in \mathbb{R}$$

である。$x = X$ として期待値を取ることで目的の不等式を得る。
f が狭義凸の場合、f と接線は一点のみを共有するので決定的でない場合は狭義の不等式が成り立つ。 $\square$

Log sum 不等式

Lemma (Log sum 不等式)

任意の $a_1, \dots, a_m, b_1, \dots, b_m \in \mathbb{R}_{>0}$ について

$$\sum_k a_k \log \frac{a_k}{b_k} \geq \left(\sum_k a_k\right) \log \frac{\left(\sum_k a_k\right)}{\left(\sum_k b_k\right)}.$$

等号は $\frac{a_k}{b_k}$ が一定の場合のみ成り立つ。

Proof.

$$\begin{aligned}
\sum_k a_k \log \frac{a_k}{b_k} &= \left(\sum_k b_k\right) \sum_k \frac{b_k}{\sum_{k'} b_{k'}} \frac{a_k}{b_k} \log \frac{a_k}{b_k} \\
&\geq \left(\sum_k b_k\right) \left(\sum_k \frac{b_k}{\sum_{k'} b_{k'}} \frac{a_k}{b_k}\right) \log \left(\sum_k \frac{b_k}{\sum_{k'} b_{k'}} \frac{a_k}{b_k}\right) \\
&\qquad \text{(}x \log x\text{ の凸性とイェンセンの不等式)} \\
&= \left(\sum_k a_k\right) \log \frac{\left(\sum_k a_k\right)}{\left(\sum_k b_k\right)}.
\end{aligned}$$

□

KL-divergence の凸性

Lemma

任意の $p, p', q, q' \in \mathcal{P}$ と $\lambda \in [0,1]$ について

$$D(\lambda q + (1-\lambda)q' \,\|\, \lambda p + (1-\lambda)p') \le \lambda D(q \,\|\, p) + (1-\lambda) D(q' \,\|\, p').$$

Proof.

$$\begin{aligned}
&(\lambda q_k + (1-\lambda)q'_k) \log \frac{\lambda q_k + (1-\lambda)q'_k}{\lambda p_k + (1-\lambda)p'_k} \\
&\le \lambda q_k \log \frac{\lambda q_k}{\lambda p_k} + (1-\lambda) q_k \log \frac{(1-\lambda)q_k}{(1-\lambda)p_k} \qquad \text{(Log sum 不等式)} \\
&= \lambda q_k \log \frac{q_k}{p_k} + (1-\lambda) q_k \log \frac{q_k}{p_k}.
\end{aligned}$$

両辺を k について和を取ると目的の不等式を得る。 □

$p' = p$ や $q' = q$ の場合を考えると、

$$\begin{aligned}
D(\lambda q + (1-\lambda)q' \,\|\, p) &\le \lambda D(q \,\|\, p) + (1-\lambda) D(q' \,\|\, p) \\
D(q \,\|\, \lambda p + (1-\lambda)p') &\le \lambda D(q \,\|\, p) + (1-\lambda) D(q \,\|\, p').
\end{aligned}$$

エントロピー

Definition (エントロピー)

$H\colon \mathcal{P} \to \mathbb{R}$ を以下で定義する

$$H(p) := -\sum_{k=1}^{m} p_k \log p_k.$$

ただし、$0 \log 0 = 0$ とする。$\log$ の底は m とすることも多い。

q を $\{1, 2, \ldots, m\}$ 上の一様分布とすると、$H(p) = D\,(p \,\|\, q) - \log m$ である。

多項係数とエントロピー

Lemma

$N, N_1, \ldots, N_m \in \mathbb{Z}_{\geq 0}$ で $N_1 + \cdots + N_m = N$ のとき、

$$\frac{1}{(N+1)^m} \mathrm{e}^{NH(p)} \leq \binom{N}{N_1\, N_2\, \ldots, N_m} \leq \mathrm{e}^{NH(p)}$$

ただし $k = 1, 2, \ldots, m$ について $p_k := N_k/N$ とする。

Proof.

$$\frac{1}{(N+1)^m} \leq \binom{N}{N_1\, N_2\, \ldots, N_m} \prod_{k=1}^{m} (p_k)^{Np_k} \leq 1$$

$$\iff \frac{1}{(N+1)^m} \leq \binom{N}{N_1\, N_2\, \ldots, N_m} \mathrm{e}^{-NH(p)} \leq 1.$$

□

データ圧縮とエントロピー

データ圧縮の概要

- データ $\mathbf{x} \in A^N$ が i.i.d でそれぞれが確率分布 $p \in \mathcal{P}$ に従っていると仮定する。
- サノフの定理よりデータの経験分布 $\hat{P}_{\mathbf{x}}$ は高い確率で真の分布 p に近い。
- よってあり得る $\mathbf{x}$ の数は大体 $\binom{N}{p_1N\, p_2N \cdots p_mN} \approx \mathrm{e}^{NH(p)}$ である。
- よって長さ $\log_m \mathrm{e}^{NH(p)} = NH(p)/(\log m)$ の A の列に圧縮できる。ここで $H(p)/(\log m)$ はエントロピーの定義の log の底を m にしたものに等しい。

確率論振り返り

- 確率空間 (Ω, P). 完全加法性。
- 大数の法則、クラメールの定理。
- 中心極限定理
- サノフの定理

課題

- 二元エントロピー関数
  $h(p) := -p\log_2 p - (1-p)\log_2(1-p) \quad \text{for } p \in [0,1]$ のグラフを描け。凸性に気をつけること。